PROBABILIDAD Y ESTADÍSTICA II

TEMA 3.2 MODELOS DE COLAS BÁSICOS

CONTENIDOS

1. Introducción a las colas poissonianas.
2. Modelo de colas poissoniano con un servidor *M/M/1*
3. Modelo con un servidor y capacidad finita *M/M/1/K*
4. Modelo con varios servidores *M/M/c*. Fórmula *C* de Erlang
5. Modelo con infinitos servidores *M/M/*∞

1. Introducción a los modelos de colas poissonianos

Las colas poissonianas (o exponenciales o markovianas) son modelos del tipo *M/M*, con llegadas de Poisson y servicio exponencial, que son las más estudiadas analíticamente.

Las llegadas de clientes y su servicio demandado son completamente aleatorios en el sentido de que la evolución del sistema depende sólo de su estado actual, y no de su pasado.

Los procesos de nacimiento y muerte introducidos sirven para describir muchos modelos de colas. Asociaremos el término nacimiento con la llegada de un cliente al sistema y el término muerte con la salida de un cliente del sistema después de completado su servicio. El número de clientes en el sistema en el instante t, $N(t)$, indica el estado del mismo.

Estudiaremos el comportamiento de las probabilidades $p_n(t)$ en el límite $\pi_n = \lim_{t\to\infty} p_n(t)$, que indica la proporción de tiempo que el sistema permanece con n clientes.

La solución de equilibrio (tema 11) se obtenía de las ecuaciones que igualaban las tasas de entrada y salida de cada estado, dando lugar a

$$\pi_0 = \frac{1}{1 + \sum_{n=1}^{\infty} \frac{\lambda_{n-1} \cdots \lambda_0}{\mu_n \cdots \mu_1}} \qquad \pi_n = \frac{\lambda_{n-1} \cdots \lambda_0}{\mu_n \cdots \mu_1} \pi_0, \quad n = 1, 2, 3... \qquad \textbf{(10.1)}$$

Para que exista dicha solución de equilibrio se debe satisfacer

$$S_1 = 1 + \sum_{n=1}^{\infty} \frac{\lambda_{n-1} \cdots \lambda_0}{\mu_n \cdots \mu_1} = \frac{1}{\pi_0} < \infty \qquad S_2 = \sum_{n=0}^{\infty} (1/(\lambda_n \prod_{i=0}^{n-1} \frac{\lambda_i}{\mu_{i+1}})) = \infty$$

que ocurre si existe un n_0 tal que $\forall\, n > n_0$, $\lambda_n / \mu_n < 1$.

Por tanto, con los diversos λ_n, μ_n que se tendrán dependiendo del modelo en estudio, las ecuaciones de S_1 y S_2 servirán para buscar las condiciones bajo las que existe solución de equilibrio π_n, mientras que con las ecuaciones de π_0 y π_n obtendremos dicha solución.

2. Modelo de colas poissoniano con un servidor *M/M/1*

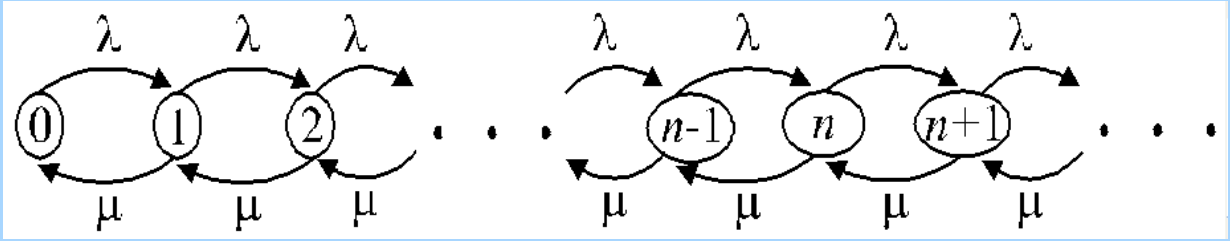
En este modelo se dispone sólo de un canal para dar servicio, las llegadas siguen un proceso de Poisson y la distribución del tiempo de servicio es exponencial.

Así, las tasas de nacimiento y muerte no dependen del número de clientes en el sistema y

$$\lambda_n = \lambda, \quad n = 0,1,2,... \qquad \mu_n = \mu, \quad n = 1,2,3,...$$

La capacidad del sistema es ilimitada y la disciplina de la cola es FIFO.

La siguiente figura representa el diagrama de transición



que conduce al sistema de ecuaciones en equilibrio

Estado	Tasa entrada	=	Tasa salida
0	$\mu\pi_1$	=	$\lambda\pi_0$
$n \geq 1$	$\lambda\pi_{n-1} + \mu\pi_{n+1}$	=	$(\lambda+\mu)\pi_n$

$$\begin{aligned}
\pi_0\lambda &= \pi_1\mu \\
\pi_1(\lambda+\mu) &= \pi_0\lambda + \pi_2\mu \\
\pi_2(\lambda+\mu) &= \pi_1\lambda + \pi_3\mu \\
&\cdots \\
\pi_i(\lambda+\mu) &= \pi_{i-1}\lambda + \pi_{i+1}\mu \\
&\cdots \\
\sum_{i=0}^{\infty} \pi_i &= 1
\end{aligned}$$

$$\begin{aligned}
\pi_0\lambda &= \pi_1\mu \\
\pi_1\lambda &= \pi_2\mu \\
\pi_2\lambda &= \pi_3\mu \\
&\cdots \\
\pi_i\lambda &= \pi_{i+1}\mu \\
&\cdots \\
\sum_{i=0}^{\infty} \pi_i &= 1
\end{aligned}$$

$$\begin{aligned}
\pi_1 &= \rho\pi_0 \\
\pi_2 &= \rho\pi_1 = \rho^2\pi_0 \\
\pi_3 &= \rho\pi_2 = \rho^3\pi_0 \\
&\cdots \\
\pi_i &= \rho\pi_{i-1} = \rho^i\pi_0 \\
&\cdots \\
\sum_{i=0}^{\infty} \pi_i &= 1
\end{aligned}$$

Sustituyendo las expresiones de los π_i en la última ecuación y despejando π_0 obtenemos (teniendo en cuenta que el factor de utilización es $\rho = r = \lambda/\mu$):

$$\pi_0 = \frac{1}{\sum_{n=0}^{\infty} \rho^n} = \frac{1}{1/(1-\rho)} = 1 - \rho$$

$$\pi_n = (1-\rho)\rho^n, \qquad n = 0, 1, 2, 3...$$

que corresponde a una distribución geométrica de parámetro 1- ρ.

$$S_1 = \sum_{n=0}^{\infty} \rho^n < \infty \qquad S_2 = \frac{1}{\lambda} \sum_{n=0}^{\infty} \frac{1}{\rho^n} = \infty$$

La serie de S_1 converge si y sólo si $\rho < 1$. La segunda condición (S_2) se satisface si $\rho \leq 1$.

Luego, la condición necesaria y suficiente para que un modelo *M/M/1* tenga solución de equilibrio, es que $\rho<1$, que es la condición de estabilidad.

Por tanto, la probabilidad de que el canal esté ocupado es

P(canal esté ocupado) = 1 - π_0 = 1 - (1 - ρ) = ρ

La probabilidad de encontrar al menos n de clientes en el sistema

$$P(N \geq n) = \sum_{k=n}^{\infty} \pi_k = \sum_{k=n}^{\infty} (1-\rho)\rho^k = (1-\rho)\rho^n \sum_{k=n}^{\infty} \rho^{k-n}$$

$$= (1-\rho)\rho^n \sum_{k=0}^{\infty} \rho^k = \frac{(1-\rho)\rho^n}{1-\rho} = \rho^n.$$

Medidas de rendimiento

Comenzando por el número medio de clientes en el sistema, L, y en la cola, L_q. Se tiene

$$\begin{aligned} L = E(N) &= \sum_{n=0}^{\infty} n\pi_n = \sum_{n=0}^{\infty} n(1-\rho)\rho^n = (1-\rho)\sum_{n=0}^{\infty} n\rho^n \\ &= (1-\rho)\rho\sum_{n=1}^{\infty} n\rho^{n-1} = (1-\rho)\rho\frac{d}{d\rho}\left(\sum_{n=0}^{\infty}\rho^n\right) \\ &= (1-\rho)\rho\frac{d}{d\rho}\left(\frac{1}{1-\rho}\right) = \frac{(1-\rho)\rho}{(1-\rho)^2} = \frac{\rho}{1-\rho}. \end{aligned}$$

La sexta igualdad se debe a que las operaciones de suma y diferenciación pueden intercambiarse cuando las funciones implicadas se comportan lo suficientemente bien.

Otra expresión equivalente, en función de λ y μ, es $L = \dfrac{\lambda}{\mu - \lambda}$ ya que $\rho = \lambda/\mu$.

L también podía haberse deducido directamente por tener *N* distribución geométrica. Nótese que *L*, como función de ρ, tiene una asíntota vertical en ρ=1, lo que indica el dramático comportamiento del número medio de clientes en el sistema según nos acercamos hacia la violación de la condición de estabilidad.

Aparte de la media que acabamos de calcular de la variable *N*, podemos obtener su varianza, a partir de la distribución geométrica

$$\sigma_N^2 = \sum_{n=0}^{\infty} (n - L)^2 \pi_n = \frac{\rho}{(1-\rho)^2}$$

Calculamos el número medio de clientes en la cola L_q mediante

$$L_q = E(N_q) = 0\pi_0 + \sum_{n=1}^{\infty} (n-1)\pi_n = \sum_{n=1}^{\infty} n\pi_n - \sum_{n=1}^{\infty} \pi_n$$

$$= L - (1 - \pi_0) = L - \rho = \frac{\rho^2}{1-\rho},$$

que en términos de λ y μ es $$L_q = \frac{\lambda^2}{\mu(\mu - \lambda)}$$

Nótese que la igualdad $L_q = L - (1 - \pi_0)$ es general para cualquier cola con un servidor y dando servicio de uno en uno, ya que para obtenerla no se ha utilizado el tipo de distribuciones de los tiempos entre llegadas o de servicio.

Otra relación entre L y L_q es $$L_q = \frac{\rho}{1-\rho}\rho = L\rho = L\frac{\lambda}{\mu} \Longrightarrow \frac{L_q}{\lambda} = \frac{L}{\mu}$$

Recordemos que siempre $N = N_q + N_s$. Pero en el modelo que estamos tratando, si $N \geq 1$, entonces $N = N_q + 1$, mientras que en general $L \neq L_q+1$, ya que L y L_q son medias y hay momentos en los que el servidor está desocupado.
Además, por las fórmulas de Little el número medio de clientes en el servidor $L_s = \lambda W_s = \rho$.

Calculamos el tamaño esperado de la cola cuando hay cola, denotado como $L_q' = E(N_q \mid N_q>0)$. Como la probabilidad condicionada de n clientes en el sistema dado que la cola no está vacía,

π_n' = $P(n$ clientes en el sistema $\mid n \geq 2)$ =
$P(n$ clientes en el sistema , $n \geq 2)/P(n \geq 2) = \pi_n / (1- \pi_0- \pi_1) = \pi_n / \rho^2$,

para $n \geq 2$, se llega a

$$L_q' = E(N_q \mid N_q > 0) = \sum_{n=2}^{\infty}(n-1)\pi_n' = \sum_{n=2}^{\infty} n\frac{\pi_n}{\rho^2} - \sum_{n=2}^{\infty}\frac{\pi_n}{\rho^2}$$
$$= \frac{(L - \pi_1) - (1 - \pi_0 - \pi_1)}{\rho^2}$$
$$= \frac{(\rho/(1-\rho)) - (1-\pi_0)}{\rho^2} = \frac{1}{1-\rho}.$$

En general, dadas dos v.a. X e Y, se verifica que $E(X) = E_Y[E(X \mid Y)]$. Esto nos ofrece un camino alternativo para obtener L_q' a partir de L_q

$$L_q = E(N_q) = E(N_q \mid N_q = 0)P(N_q = 0) + E(N_q \mid N_q > 0)P(N_q > 0)$$
$$= E(N_q \mid N_q > 0)P(N_q > 0),$$

de donde $$E(N_q \mid N_q > 0) = \frac{L_q}{P(N_q > 0)} = \frac{L_q}{P(N \geq 2)} = \frac{L_q}{\rho^2} = \frac{1}{1-\rho}.$$

Ejemplo. En un pequeño servidor el tiempo de procesamiento por trabajo se distribuye exponencialmente con un tiempo medio de 3 minutos. Los trabajos llegan aleatoriamente cada 4 minutos en media. Los trabajos se procesan con la disciplina FIFO.

Calculemos primero las tasas de nacimiento y muerte: $\lambda = 1/4$ trabajos/minuto, $\mu = 1/3$ trabajos/minuto.

Luego, el factor de utilización es $\rho=\lambda/\mu=3/4=0.75<1$, que indica que existe solución de equilibrio.

Si lo que nos preocupa es la probabilidad de que entre la llegada de dos trabajos consecutivos transcurran más de, digamos, 15 minutos, podemos obtenerla recordando que el tiempo entre llegadas consecutivas es $\zeta \sim \text{Exp}(\lambda=1/4)$.

Por tanto, dicha probabilidad es $P(\zeta > 15) = e^{-0.25\times15} = 0.0235$.

Las siguientes probabilidades pueden también ser de interés

$$P(\text{tener que esperar}) = 1 - \pi_0 = \rho = 0.75$$
$$P(\text{encontrar cola}) = P(N \geq 2) = \rho^2 = 0.5625$$

Es decir, sólo el 25% de los trabajos pasarán inmediatamente a recibir servicio y el 56.25% encontrarán cola al llegar.

Por otra parte, el número medio de trabajos en el sistema, L, y en cola, L_q, es

$$L = \rho/(1-\rho) = 0.75/0.25 = 3 \text{ trabajos y } L_q=\rho\, L = 2.25 \text{ trabajos.}$$

La varianza de la v.a. N es $\sigma_N^2 = 12$, por tanto, podemos decir que N es una v.a. discreta con valores 0,1,2,... y probabilidades respectivas π_0, π_1, π_2,..., media 3 y varianza 12.

El número medio de trabajos en cola, cuando hay cola, es $L_q'=1/(1-\rho)= 4$ trabajos.

El dramático comportamiento de L según $\rho \to 1$ puede observarse cuando aumenta ρ, bien porque aumenta la tasa de llegadas o bien porque disminuye la de servicio.

Así, si aumentase un 25% siendo 18.75 trabajos/hora, elevaría ρ hasta 0.9375 y en consecuencia L = 15 trabajos, que es cinco veces la que se tenía anteriormente.

Tiempos de espera

Sólo resta estudiar los tiempos de espera del modelo *M/M/1*. Obtendremos no sólo las medias W y W_q, sino también las distribuciones de probabilidad de las v.a. w y q.

Las medias se calculan fácilmente por las fórmulas de Little:

$$W = E(w) = \frac{L}{\lambda} = \frac{\rho}{\lambda(1-\rho)} = \frac{E(s)}{1-\rho}$$

$$W_q = E(q) = \frac{L_q}{\lambda} = \frac{\rho^2}{\lambda(1-\rho)} = \frac{\rho E(s)}{1-\rho}$$

Nótese que como dijimos para L, W tiene también un comportamiento dramático según ρ tiende a 1.

Ahora queremos calcular el tiempo medio de espera en cola para aquellos clientes que deben esperar.

Como

$$\begin{aligned} W_q = E(q) &= E(q \mid q = 0)P(q = 0) + E(q \mid q > 0)P(q > 0) \\ &= 0(1 - \rho) + E(q \mid q > 0)\rho, \end{aligned}$$

se tiene

$$E(q \mid q > 0) = \frac{W_q}{\rho} = \frac{E(s)}{1 - \rho} = W$$

Esta cantidad interesa porque un tiempo medio de espera aceptable puede deberse a que muchos clientes no tienen que esperar, pero los que esperan lo hacen durante mucho tiempo.

Como $W = W_q + E(s)$, tenemos que $E(q \mid q > 0) = W_q + E(s)$, lo que indica que, en media, los clientes que tienen que esperar en cola esperan más que lo que un cliente medio espera, ya que espera un tiempo medio de servicio, $E(s)$, más.

Para hallar la distribución de la variable aleatoria q, nótese primero que tiene un punto ($t = 0$) con probabilidad positiva: $P(q = 0) = P(N = 0) = 1 - \rho$.

Por otra parte, si al llegar el cliente encuentra n personas en el sistema (una de ellas en el servidor), tendrá que esperar a que todas se sirvan. Así, el tiempo de espera en cola es la suma de las variables aleatorias "tiempo de servicio del cliente i ", i =1,...,n,

$$q = s_1 + \ldots + s_n$$

en donde s_i son independientes e idénticamente distribuidas según una exponencial de parámetro μ.

Debemos recordar que por la pérdida de memoria de la distribución exponencial, no hace falta tener en cuenta el tiempo de servicio ya consumido por el cliente que actualmente está sirviéndose.

Por la reproductividad de la distribución gamma, $q \mid N=n$ sigue una distribución gamma de parámetros $p = n$, a = μ (en este caso es Erlang, al ser p entero).

Por el teorema de la probabilidad total se tiene

$$
\begin{aligned}
P(0 < q \leq t) &= \sum_{n=1}^{\infty} P(q \leq t \mid N = n) P(N = n) \\
&= \sum_{n=1}^{\infty} \int_0^t \mu e^{-\mu x} \frac{(\mu x)^{n-1}}{(n-1)!} \rho^n (1-\rho) dx \\
&= \int_0^t (1-\rho) e^{-\mu x} \sum_{n=1}^{\infty} \frac{(\mu \rho x)^{n-1}}{(n-1)!} \rho \mu dx \\
&= \int_0^t \mu \rho (1-\rho) e^{-\mu x} e^{\mu \rho x} dx = \rho (1 - e^{-\mu(1-\rho)t}) = \rho (1 - e^{-t/W}).
\end{aligned}
$$

Luego, la función de distribución de *q* es

$$
F_q(t) = P(q = 0) + P(0 < q \leq t) = 1 - \rho + \rho (1 - e^{-t/W}) = 1 - \rho e^{-t/W}.
$$

Esta expresión es válida para todo *t*, aunque *q* sea discreta en el origen y continua para $t > 0$.

Análogamente, calculamos la distribución de la v.a. *w*. Si cuando llega un cliente ya hay *n* en el sistema, éste tendrá que estar en el sistema un tiempo total igual a la suma de *n* + 1 v.a.i.i.d. según una ley exponencial de parámetro μ.

Así, la distribución de *w* será una gamma de parámetros *p* = *n*+1, *a* = μ. Variando *n*, por el teorema de la probabilidad total:

$$F_w(t) = P(w \le t) = \sum_{n=0}^{\infty} P(w \le t \mid N = n) P(N = n)$$

$$= \sum_{n=0}^{\infty} \int_0^t \mu e^{-\mu x} \frac{(\mu x)^n}{n!} \rho^n (1-\rho) dx = \int_0^t \mu(1-\rho) e^{-\mu x} \sum_{n=0}^{\infty} \frac{(\mu \rho x)^n}{n!} dx$$

$$= \int_0^t \mu(1-\rho) e^{-\mu x} e^{\mu \rho x} dx = 1 - e^{-\mu(1-\rho)t} = 1 - e^{-t/W} \quad (t \ge 0).$$

Es decir, *w* sigue una distribución exponencial de parámetro μ(1 – *ρ*) = μ –λ = 1/W.

Ejemplo. Analizar los tiempos en el sistema y en la cola para el ejemplo anterior. Además, supongamos que se decide aumentar la capacidad del servidor cuando la carga de trabajo llegue a un nivel tal que el tiempo medio en el sistema alcance los 30 minutos. Determinar la tasa media de llegada de trabajos a la que ocurrirá esto. Repetir el cálculo si el criterio para aumentar la capacidad del servidor fuese que no más del 10% de los trabajos empleen más de 45 minutos en el sistema.

Del ejemplo anterior sabemos las tasas de nacimiento y muerte son $\lambda = 1/4$ trabajos/min, $\mu = 1/3$ trabajos/min, respectivamente, por lo que $\rho = \lambda / \mu = 3/4 = 0.75$.

Sabemos que $W = E(s)/(1-\rho) = 3/(1-0.75) = 12$ minutos, y $W_q = \rho W = 9$ minutos.

El tiempo medio de espera en cola de los programas que tienen que esperar es $E(q \mid q > 0) = W = 12$ minutos.

Además, al conocer las distribuciones de probabilidad de q y w podemos preguntar por distintas probabilidades de espera o permanencia en el sistema.

Por ejemplo, la probabilidad de que un trabajo tenga que esperar en la cola más de 20 minutos es

$$P(q > 20) = 1 - P(q \leq 20) = 1 - (1 - \rho\, e^{-t/W}) = \rho\, e^{-t/W} = 0.75\, e^{-20/12} = 0.142.$$

La probabilidad de permanecer en el sistema más de 20 minutos es

$$P(w > 20) = 1 - P(w \leq 20) = 1 - (1 - e^{-t/W}) = e^{-t/W} = e^{-20/12} = 0.188.$$

Luego, más del 14% de los programas estarán en la cola más de 20 minutos y casi el 20% no saldrán del sistema en menos de 20 minutos.

Según indica el enunciado, se decide aumentar μ cuando W sea 30 minutos debido a un aumento de la carga de trabajo ρ por aumentar λ. Para hallar el valor de para el que esto ocurrirá, resolvemos

$$30 = W = \frac{E(s)}{1 - \lambda E(s)} = \frac{3}{1 - 3\lambda} \Longrightarrow \lambda = 0.3 \text{ trabajos/minuto.}$$

El segundo criterio para aumentar la capacidad del servidor exige que

$$P(w > 45) \leq 0.1 \rightarrow 1 - P(w \leq 45) \leq 0.1 \rightarrow 1 - (1 - e^{-45/W}) \leq 0.1 \rightarrow$$

$$e^{-45/W} \leq 0.1 \rightarrow -45/W \leq \ln(0.1) \rightarrow W \geq 19.54$$

$$19.54 = W = \frac{E(s)}{1 - \lambda E(s)} = \frac{3}{1 - 3\lambda} \Longrightarrow \lambda = 0.282 \text{ trabajos/minuto},$$

que representa un incremento del 12.86% sobre la actual tasa de llegadas.

Ahora, los tamaños medios L y L_q pasan a ser 5.51 y 4.67 trabajos, respectivamente, suponiendo incrementos del 83.3% y 207.5%, que son menores que anteriormente (L= 3 trabajos y L_q = 2.25 trabajos).

El tiempo medio en el sistema es W = 19.5 minutos.

Ejemplo. En una compañía se acaba de montar una red local y se observa que la caída de cada componente se produce según un proceso de Poisson con tasa media de 2 por hora durante las 8 horas de trabajo diario. La compañía está considerando contratar los servicios de mantenimiento de dos candidatos.

El tiempo que emplea el primero en restaurar la red depende del problema encontrado, pero se ajusta a una distribución exponencial con una tasa media de 4 componentes por hora, con unos costes por su servicio de 30 euros por hora.

El segundo candidato actúa con un tiempo de mantenimiento exponencial con una tasa media de 6 componentes por hora, cobrando 50 euros por hora.

Encontrar el mejor candidato, sobre una base diaria, si el coste de un componente fuera de servicio es de 36 euros por hora.

Se tiene una tasa $\lambda = 2$ caídas/hora. El modelo *M/M/1* para el primer candidato verifica $\mu_1 = 4$ componentes/hora, por lo que $\rho_1 = 0.5$.

En un día, esta persona cobrará en media $0.5 \times 8 \times 30 = 120$ euros, ya que los dos primeros factores dan el número de horas que trabaja.

Además, hay que contar el coste diario por tener los componentes fuera de servicio, que se calculará como el producto del coste de cada hora (36 euros), el número medio de componentes que hay que mantener en un día (8×2) y el tiempo medio que pasa cada uno caído ($W = E(s)/(1-\rho) = (1/4)/0.5 = 0.5$ horas/componente). Es decir, 36 × 16 × 0.5 = 288 euros/día.

El coste total es 408 euros/día.

Con el segundo candidato, $\mu_2 = 6$ componentes/hora, por lo que $\rho_2 = 2/6$.

El coste por su servicio es, en media, de 2/6 × 8 × 50 = 133.33 euros/día.

El coste diario por tener componentes fuera de servicio es de 36 × 16 × 0.25 = 144 euros, suponiendo un coste total de 277.33 euros/día, más barato que con el primer candidato.

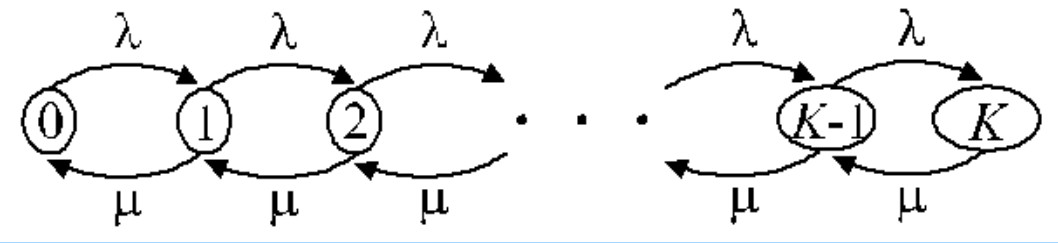
3. Modelo *M/M/1/K*: Capacidad *K* finita del sistema

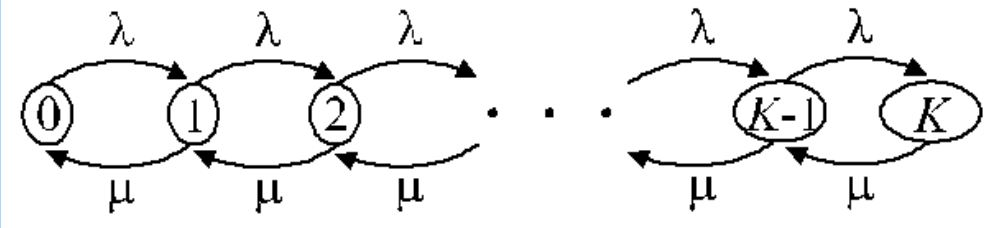
Se admite a lo sumo un número *K* de clientes en el sistema, de forma que no se permiten más entradas en el sistema si se alcanza tal cota, siendo rechazadas. Así, las tasas de nacimiento y muerte dependen del número de clientes en el sistema

$$\lambda_n = \begin{cases} \lambda, & \text{si } n = 0, 1, ..., K-1 \\ 0, & \text{si } n \geq K \end{cases}$$

$$\mu_n = \begin{cases} \mu, & \text{si } n = 1, 2, ..., K \\ 0, & \text{si } n > K \end{cases}$$

y su diagrama de transición es





El sistema de ecuaciones en equilibrio es:

Estado	Tasa entrada	=	Tasa salida
0	$\mu\pi_1$	$=$	$\lambda\pi_0$
$1 \leq n \leq K-1$	$\lambda\pi_{n-1} + \mu\pi_{n+1}$	$=$	$(\lambda+\mu)\pi_n$
K	$\lambda\pi_{K-1}$	$=$	$\mu\pi_K$

Resolviéndolo se obtiene:

$$\lambda\pi_0 = \mu\pi_1 \quad \Rightarrow \pi_1 = \frac{\lambda}{\mu}\pi_0$$

$$\lambda\pi_1 = \mu\pi_2 \quad \Rightarrow \pi_2 = \frac{\lambda}{\mu}\pi_1 = \left(\frac{\lambda}{\mu}\right)^2 \pi_0$$

$$\ldots$$

$$\lambda\pi_{K-1} = \mu\pi_K \quad \Rightarrow \pi_K = \left(\frac{\lambda}{\mu}\right)^K \pi_0$$

Sumando todas las ecuaciones, como $\sum_{n=0}^{K} \pi_n = 1$, pues $\pi_n = 0$ para $n > K$, se tiene que

$$1 = \pi_0 \sum_{n=0}^{K} \left(\frac{\lambda}{\mu}\right)^n = \pi_0 \frac{1 - \left(\frac{\lambda}{\mu}\right)^{K+1}}{1 - \frac{\lambda}{\mu}},$$

Por tanto, si $\lambda \neq \mu$,

$$\pi_n = \left(\frac{\lambda}{\mu}\right)^n \frac{1 - \frac{\lambda}{\mu}}{1 - \left(\frac{\lambda}{\mu}\right)^{K+1}}, \quad n = 0, 1, 2, ..., K.$$

Puede comprobarse usando las expresiones de S_1 y S_2 que en este modelo existe solución de equilibrio para todo λ y μ, incluso para $\lambda \geq \mu$.

El truncamiento del sistema a K clientes lo explica, pues el sistema nunca se desborda ni crece indefinidamente al rechazar a los clientes que llegan cuando está lleno (la cadena de Markov asociada es irreducible y finita y, por tanto, ergódica).

Si λ = μ, la distribución de probabilidad del número de clientes en el sistema es uniforme

$$\pi_n = \frac{1}{K+1}, \quad n = 0, 1, 2, ..., K.$$

Si eliminásemos el truncamiento, es decir $K \to \infty$, cuando λ < μ las expresiones de π_n se convierten en las obtenidas para el modelo *M/M/1*.

Medidas de rendimiento

Comenzamos con el número medio de clientes en el sistema. Para λ = μ,

$$L = E(N) = \sum_{n=0}^{K} n\pi_n = \frac{1}{K+1}\sum_{n=0}^{K} n = \frac{K(K+1)}{2(K+1)} = \frac{K}{2},$$

que ya esperábamos por ser uniforme.

Para λ ≠ μ, siendo $u = \lambda/\mu$,

$$\begin{aligned} L = E(N) &= \sum_{n=0}^{K} n\pi_n = \sum_{n=0}^{K} nu^n\pi_0 = \pi_0 u \sum_{n=1}^{K} nu^{n-1} = \pi_0 u \frac{d}{du}\left(\sum_{n=0}^{K} u^n\right) \\ &= \pi_0 u \frac{d}{du}\left(\frac{1-u^{K+1}}{1-u}\right) = \frac{u[1-(K+1)u^K + Ku^{K+1}]}{(1-u)(1-u^{K+1})} \\ &= \frac{\lambda[1-(K+1)(\lambda/\mu)^K + K(\lambda/\mu)^{K+1}]}{(\mu-\lambda)[1-(\lambda/\mu)^{K+1}]}, \end{aligned}$$

que también puede expresarse como

$$\frac{u}{1-u} - \frac{(K+1)u^{K+1}}{1-u^{K+1}}$$

donde el primer sumando es la expresión de *L* del modelo *M/M/1*. Por tanto, el número esperado de clientes en el sistema *M/M/1/K* es siempre menor que en el *M/M/1*, haciéndolo más eficiente.

Como para todo λ y μ se tiene

$$\begin{aligned} L_s = E(N_s) &= E(N_s \mid N=0)P(N=0) + E(N_s \mid N>0)P(N>0) \\ &= 0\pi_0 + 1(1-\pi_0) = 1-\pi_0, \end{aligned}$$

entonces $L_q = L - L_s = L - (1 - \pi_0)$.

En este modelo se rechaza a los clientes que llegan cuando ya hay *K* en el sistema (*K*-1 en la cola), lo que ocurre con probabilidad π_K.

Luego, la probabilidad de que al llegar un cliente entre en el sistema es $1-\pi_K$, representando la proporción de tiempo que el sistema no está saturado o la proporción de clientes que llegan que realmente entran en el sistema.

Así, la tasa media de entradas al sistema o paso a través del sistema, $\lambda_e = \boldsymbol{\lambda}$, se define como

$$\lambda_e = \lambda(1 - \pi_K).$$

La utilización verdadera del servidor, ρ, probabilidad de que el servidor esté ocupado, ya no es $u = \lambda/\mu$ y de ahí que lo hayamos etiquetado u en vez de r, sino

$$\rho = \lambda_e W_s = \frac{\lambda}{\mu}(1 - \pi_K) = 1 - \pi_0.$$

Tiempos de espera

Entendiendo por clientes en el sistema aquellos que entran en el sistema, podemos aplicar las fórmulas de Little para conseguir los tiempos medios en el sistema y en la cola

$$W = L / \lambda_e \qquad W_q = L_q / \lambda_e$$

Para obtener el tiempo medio de espera en cola para aquellos clientes que deben esperar, hacemos como en el modelo *M/M/1*,

$$E(q \mid q > 0) = W_q/\rho = W_q/(1- \pi_0).$$

El proceso de obtención de las distribuciones de los tiempos *q* y *w* es más complejo que en el modelo *M/M/1*. Como hicimos entonces, utilizaremos el teorema de la probabilidad total, pero ahora condicionando a la v.a. N_e, que cuenta el número de clientes en el sistema cuando entra un cliente en él.

Denotamos con $q_n=P(N_e=n)$, n=0,1,2,...,K-1, la probabilidad de que un cliente que entra en el sistema encuentre n clientes en él, que por el teorema de Bayes es

$$q_n = \frac{\pi_n}{1-\pi_K}, \quad n = 0, 1, 2, ..., K-1.$$

Nótese que en este modelo la entrada no es una verdadera Poisson, $p_n \neq q_n$, y $\lambda_n = \lambda$ para $n \leq K$-1 pero $\lambda_n = 0$ para $n \geq K$, a diferencia de lo que ocurría en el *M/M/1*. Así, para $t \geq 0$,

$$\begin{aligned}
F_w(t) = P(w \leq t) &= \sum_{n=0}^{K-1} P(w \leq t \mid N_e = n) P(N_e = n) \\
&= \sum_{n=0}^{K-1} \left[\int_0^t \mu e^{-\mu x} \frac{(\mu x)^n}{n!} dx \right] q_n = \sum_{n=0}^{K-1} \left[1 - \int_t^\infty \mu e^{-\mu x} \frac{(\mu x)^n}{n!} dx \right] q_n \\
&= \sum_{n=0}^{K-1} q_n - \sum_{n=0}^{K-1} q_n \left[\int_t^\infty \mu e^{-\mu x} \frac{(\mu x)^n}{n!} dx \right] = \\
&= 1 - \sum_{n=0}^{K-1} q_n \left[\sum_{i=0}^{n} \frac{(\mu t)^i e^{-\mu t}}{i!} \right] = 1 - \sum_{n=0}^{K-1} q_n F_{\mathcal{P}(\mu t)}(n),
\end{aligned}$$

donde $$F_{\mathcal{P}(\mu t)}(n) = \sum_{i=0}^{n} \frac{(\mu t)^i e^{-\mu t}}{i!} = \int_t^{\infty} \mu \frac{\mu^n x^n e^{-\mu x}}{n!} dx$$

(igualdad debida a la relación entre las distribuciones de Erlang y Poisson) es la función de distribución de Poisson de parámetro μt en el punto n, que puede obtenerse a partir de las tablas de dicha distribución.
De forma similar

$$F_q(t) = P(q = 0) + P(0 < q \leq t) = q_0 + \sum_{n=1} P(q \leq t \mid N_e = n) P(N_e = n)$$

$$= q_0 + \sum_{n=1}^{K-1} \left[\int_0^t \mu e^{-\mu x} \frac{(\mu x)^{n-1}}{(n-1)!} dx \right] q_n = q_0 + \sum_{n=0}^{K-2} q_{n+1} \left[1 - \int_t^{\infty} \mu e^{-\mu x} \frac{(\mu x)^n}{n!} dx \right]$$

$$= q_0 + \sum_{n=0}^{K-2} q_{n+1} - \sum_{n=0}^{K-2} q_{n+1} \left[\int_t^{\infty} \mu e^{-\mu x} \frac{(\mu x)^n}{n!} dx \right] = 1 - \sum_{n=0}^{K-2} q_{n+1} \left[\sum_{i=0}^{n} \frac{(\mu t)^i e^{-\mu t}}{i!} \right]$$

$$= 1 - \sum_{n=0}^{K-2} q_{n+1} F_{\mathcal{P}(\mu t)}(n)$$

Ejemplo. Un servidor de Internet tiene una velocidad de transmisión de 1600 caracteres por segundo para atender las peticiones que le llegan, que lo hacen según un proceso de Poisson con una velocidad media de 300 peticiones por minuto.
La longitud de cada petición puede aproximarse a una distribución exponencial de media 280 caracteres por petición.

Calcular las principales medidas estadísticas de eficiencia del sistema suponiendo que:

a) Se dispone de un número ilimitado de buffers; y

b) El número de buffers es 14. ¿Son suficientes 14 buffers para que la probabilidad de que el sistema esté completo no supere el 1%? En caso negativo, encontrar el número de buffers necesarios.

En a) el modelo es *M/M/1* con λ=300 peticiones/minuto, es decir, 5 peticiones/segundo y μ=(1600 caracteres/segundo)/(280 caracteres/petición)=5.714 peticiones/segundo.
Luego, ρ = 5/5.714 = 0.875.

En b) se propone un sistema *M/M/1/15*, pues se permiten 14 peticiones encoladas en los buffers más la petición siendo transmitida.

$$\rho = 1 - \pi_0 = 1 - \left(\frac{\lambda}{\mu}\right)^0 \frac{1 - \frac{\lambda}{\mu}}{1 - \left(\frac{\lambda}{\mu}\right)^{K+1}} = 1 - \frac{1 - \frac{5}{5.714}}{1 - \left(\frac{5}{5.714}\right)^{16}} = 0.858$$

El número medio de clientes en el sistema y en la cola son:

$$L = \frac{u}{1-u} - \frac{(K+1)u^{K+1}}{1-u^{K+1}} = \frac{\frac{5}{5.714}}{1 - \frac{5}{5.714}} - \frac{16\left(\frac{5}{5.714}\right)^{16}}{1 - \left(\frac{5}{5.714}\right)^{16}} = 4.86 \text{ peticiones}$$

$$L_q = L - L_s = L - (1 - \pi_0) = L - \rho = 4.86 - 0.858 = 4.002 \text{ peticiones}$$

Los tiempos medios en el sistema y en la cola son:

$$W = L/\lambda_e = \frac{L}{\lambda(1-\pi_K)} = \frac{4.86}{5(1-0.1917)} = 0.991 \text{ segundos}$$

$$W_q = L_q/\lambda_e = \frac{L_q}{\lambda(1-\pi_K)} = \frac{4.002}{5(1-0.01927)} = 0.816 \text{ segundos}$$

siendo

$$\pi_K = \pi_{15} = \left(\frac{5}{5.714}\right)^{15} \frac{1-\frac{5}{5.714}}{1-\left(\frac{5}{5.714}\right)^{16}} = 0.01917$$

La siguiente tabla recoge compara los resultados obtenidos en el sistema *M/M/1/15* con el sistema *M/M/1*:

	$M/M/1$	$M/M/1/15$
λ_e	5 peticiones/segundo	4.904 peticiones/segundo
π_0	0.125	0.142
ρ	0.875	0.858
L	7 peticiones	4.86 peticiones
L_q	6.125 peticiones	4.002 peticiones
W	1.4 segundos	0.991 segundos
W_q	1.225 segundos	0.816 segundos

Se observa una mayor eficiencia del modelo *M/M/1/15*, pero a costa de rechazar un $100\pi_{15} = 1.91\%$ de las peticiones, que deberán intentarlo más tarde o simplemente se perderán, con las consecuentes pérdidas asociadas.

Hemos visto que con 14 buffers la probabilidad de que el sistema esté lleno es algo mayor que 0.01, pues es $\pi_{15} = 0.0191$. Se puede comprobar que hacen falta 19 buffers, ya que $\pi_{20} = 0.0092$ y $\pi_{19} = 0.0106$.

Ejemplo. Un mecánico tiene un taller en el que sólo caben 4 coches. Los coches llegan según un proceso de Poisson de tasa 3 coches por día.

El mecánico tarda en arreglar un coche un tiempo distribuido exponencialmente de media 2 días, si hay 2 o menos coches en total.

Cuando hay 3 ó 4 coches, llama a un familiar para que le ayude (ambos arreglan juntos los coches), reduciendo el tiempo medio a 1 día.

Encontrar la proporción de tiempo que ambos están ocupados y la proporción de tiempo que trabaja el mecánico.

En este sistema hay 5 estados: N =0,1,2,3,4 coches en el taller, pues la capacidad es 4.
La tasa de nacimiento es $\lambda_n=\lambda=3$ coches diarios, n = 0,1,2,3. Sin embargo, la tasa de muerte depende del número de coches en el taller: $\mu_1= \mu_2= 0.5$, $\mu_3= \mu_4= 1$ coches diarios.

Éste es un ejemplo en el que el *servicio es dependiente del estado*. Las ecuaciones de equilibrio son entonces

$$
\begin{aligned}
0.5\pi_1 &= 3\pi_0 \\
3.5\pi_1 &= 0.5\pi_2 + 3\pi_0 \\
3.5\pi_2 &= 3\pi_1 + \pi_3 \\
4\pi_3 &= 3\pi_2 + \pi_4 \\
3\pi_3 &= \pi_4 \\
1 &= \pi_0 + \pi_1 + \pi_2 + \pi_3 + \pi_4
\end{aligned}
$$

cuya solución es: π_0=1/475, π_1=6/475, π_2=36/475, π_3=108/475, π_4=324/475.

Así, la probabilidad de que ambos estén ocupados es la probabilidad de que trabaje el familiar, que es $\pi_3 + \pi_4 \simeq 0.9095$. Sin embargo, el mecánico trabaja $1 - \pi_0 \simeq 0.9979$ del tiempo.

π_0=0.0021 será la proporción de tiempo en que los dos trabajadores están ociosos.

Obsérvese lo alta que es la probabilidad de rechazar los coches que llegan, π_4.

4. Modelo *M/M/c*: *c* servidores en paralelo

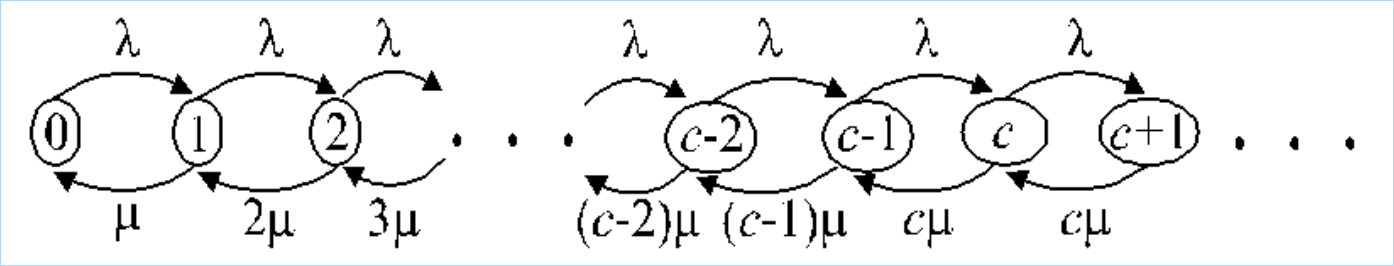
Se dispone de *c* servidores paralelos idénticos, cada uno de los cuales sirve a una tasa de μ clientes por unidad de tiempo.

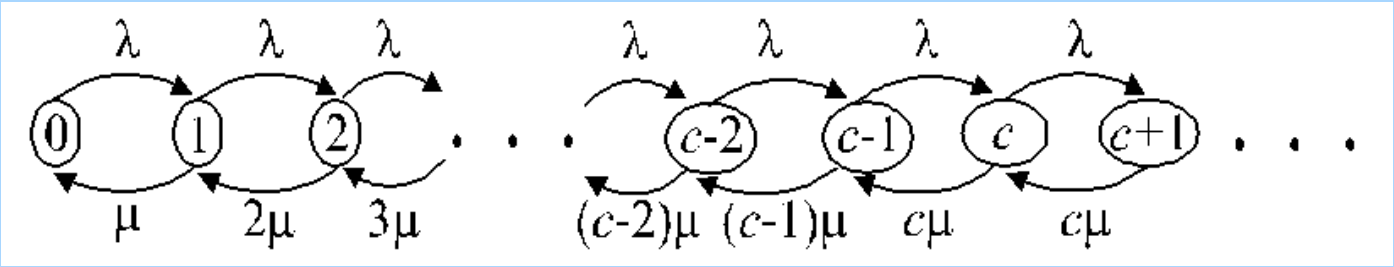
Luego, si los *c* están utilizándose, la tasa media de salida del servicio es *c*μ. Cuando hay *n* < *c* clientes en el sistema, sólo trabajan *n* servidores y, por tanto, la tasa de servicio es *n*μ. Es decir, las tasas de nacimiento y muerte son

$$\lambda_n = \lambda, \qquad n = 0, 1, 2, \ldots$$

$$\mu_n = \begin{cases} n\mu, & \text{si } n = 1, 2, \ldots, c \\ c\mu, & \text{si } n \geq c \end{cases}$$

y su diagrama de transición es





El sistema de ecuaciones en equilibrio es:

Estado	Tasa entrada	=	Tasa salida
0	$\mu\pi_1$	$=$	$\lambda\pi_0$
$1 \leq n \leq c-1$	$\lambda\pi_{n-1} + (n+1)\mu\pi_{n+1}$	$=$	$(\lambda + n\mu)\pi_n$
$n \geq c$	$\lambda\pi_{n-1} + c\mu\pi_{n+1}$	$=$	$(\lambda + c\mu)\pi_n$

El proceso para alcanzar la solución del sistema es el siguiente:

$$
\begin{aligned}
\pi_0\lambda &= \pi_1\mu \\
\pi_1(\lambda+\mu) &= \pi_0\lambda + \pi_2 2\mu \\
\pi_2(\lambda+\mu) &= \pi_1\lambda + \pi_3 3\mu \\
&\dots \\
\pi_{c-1}(\lambda+(c-1)\mu) &= \pi_{c-2}\lambda + \pi_c c\mu \\
\pi_c(\lambda+c\mu) &= \pi_{c-1}\lambda + \pi_{c+1} c\mu \\
\pi_{c+1}(\lambda+c\mu) &= \pi_c\lambda + \pi_{c+2} c\mu \\
&\dots \\
\sum_{i=0}^{\infty}\pi_i &= 1
\end{aligned}
$$

$$
\begin{aligned}
\pi_0\lambda &= \pi_1\mu \\
\pi_1\lambda &= \pi_2 2\mu \\
\pi_2\lambda &= \pi_3 3\mu \\
&\dots \\
\pi_{c-1}\lambda &= \pi_c c\mu \\
\pi_c\lambda &= \pi_{c+1} c\mu \\
\pi_{c+1}\lambda &= \pi_{c+2} c\mu \\
&\dots \\
\sum_{i=0}^{\infty}\pi_i &= 1
\end{aligned}
$$

$$
\begin{aligned}
\pi_1 &= \frac{\lambda}{\mu}\pi_0 \\
\pi_2 &= \left(\frac{\lambda}{\mu}\right)^2 \frac{1}{2!}\pi_0 \\
&\dots \\
\pi_{c-1} &= \left(\frac{\lambda}{\mu}\right)^{c-1} \frac{1}{(c-1)!}\pi_0 \\
\pi_c &= \left(\frac{\lambda}{\mu}\right)^{c} \frac{1}{c!}\pi_0 \\
\pi_{c+1} &= \left(\frac{\lambda}{\mu}\right)^{c+1} \frac{1}{c!c}\pi_0 \\
&\dots \\
\sum_{i=0}^{\infty}\pi_i &= 1
\end{aligned}
$$

Obteniendo finalmente ($r = \lambda/\mu$ es la intensidad de tráfico):

$$\pi_0 = \left[\sum_{n=0}^{c-1} \frac{r^n}{n!} + \frac{r^c}{c!}\sum_{n=c}^{\infty}(r/c)^{n-c}\right]^{-1} = \left[\sum_{n=0}^{c-1} \frac{r^n}{n!} + \frac{r^c}{c!(1-\frac{\lambda}{c\mu})}\right]^{-1}$$

y

$$\pi_n = \begin{cases} \dfrac{r^n}{n!}\pi_0, & \text{si } n = 0, 1, ..., c \\ \dfrac{r^n}{c!c^{n-c}}\pi_0, & \text{si } n \geq c \end{cases}$$

Para obtener π_0 hemos impuesto que el factor de utilización $\rho = \lambda/(c\mu) < 1$, que es la condición de estabilidad.

Una probabilidad de interés en este modelo es la probabilidad de tener que esperar en la cola (todos los servidores están ocupados), es decir, $P(N \geq c)$, que se denota como $C(c, r)$, llamada fórmula C de Erlang:

$$C(c,r) = P(N \geq c) = 1 - P(N < c) = 1 - \sum_{n=0}^{c-1} \pi_n = 1 - \sum_{n=0}^{c-1} \frac{r^n}{n!}\pi_0$$

$$= 1 - \pi_0 \left(\frac{1}{\pi_0} - \frac{r^c}{c!(1 - \frac{\lambda}{c\mu})} \right) = \pi_0 \frac{r^c}{c!(1-\rho)} = \frac{\pi_c}{1-\rho}$$

$$= \frac{r^c/c!}{(1-\rho)\left[\sum_{n=0}^{c-1} \frac{r^n}{n!} + \frac{r^c}{c!(1-\rho)}\right]},$$

Normalmente, se deja al software (por ejemplo, WinQSB) que calcule los valores $C(c,r)$, si bien tradicionalmente se obtenían de forma aproximada a partir de su representación gráfica (Allen, 1978). Hoy en día es muy sencillo programar estas fórmulas.

Medidas de rendimiento

Comenzamos calculando L_q, por ser más sencillo que L

$$L_q = E(N_q) = \sum_{n=c}^{\infty}(n-c)\pi_n = \sum_{n=0}^{\infty} n\pi_{n+c} = \sum_{n=0}^{\infty} n\frac{r^{n+c}}{c!c^n}\pi_0 = \sum_{n=0}^{\infty} n\frac{r^c r^n}{c!c^n}\pi_0 =$$
$$= \sum_{n=0}^{\infty} n\frac{r^c\rho^n}{c!}\pi_0 = \pi_0\frac{r^c}{c!}\sum_{n=0}^{\infty} n\rho^n = \pi_0\frac{r^c}{c!}\frac{\rho}{(1-\rho)^2} = C(c,r)\frac{\rho}{1-\rho},$$

Empleando las fórmulas de Little llegamos a:

$$W_q = \frac{L_q}{\lambda} = \frac{C(c,r)}{c\mu(1-\rho)}.$$

El tiempo medio de espera en cola para aquellos clientes que deben esperar:

$$E(q \mid q>0) = \frac{W_q}{P(q>0)} = \frac{W_q}{C(c,r)} = \frac{1}{c\mu(1-\rho)}.$$

A partir de aquí, podemos conseguir expresiones para *W* y *L* :

$$W = W_q + W_s = \frac{1}{\mu}\left(1 + \frac{C(c,r)}{c(1-\rho)}\right)$$

$$L = \lambda W = c\rho + C(c,r)\frac{\rho}{1-\rho}$$

Obtengamos las distribuciones de las v.a. *q* y *w*.

Para *q*, debemos tener en cuenta que el cliente que no espera en cola (*q* = 0) es el que al llegar encuentra en el sistema *N* = *n* < *c* clientes. En caso contrario, con *n* ≥ *c*, la longitud de la cola es *n* - *c* y el cliente tendrá que esperar a que se sirvan *n*-*c*+1 clientes (el que está siendo servido también cuenta).

De este modo, su tiempo en cola es $q = s_1 + \ldots + s_{n-c+1}$, con s_i v.a.i.i.d. según Exp(*c*μ), que conduce a que *q* siga una distribución gamma de parámetros *p*= *n*-*c*+1 y *a* = *c*μ.

Por lo tanto, para $t \geq 0$

$$F_q(t) = P(q = 0) + P(0 < q \leq t) = 1 - C(c,r) + \sum_{n=c}^{\infty} P(q \leq t \mid N = n)\pi_n$$

$$= 1 - C(c,r) + \sum_{n=c}^{\infty} \left[\int_0^t c\mu e^{-c\mu x} \frac{(c\mu x)^{n-c}}{(n-c)!} dx \right] \frac{r^n}{c! c^{n-c}} \pi_0$$

$$= 1 - C(c,r) + \frac{\pi_0 r^c}{(c-1)!} \int_0^t \mu e^{-c\mu x} \left(\sum_{n=c}^{\infty} \frac{(r\mu x)^{n-c}}{(n-c)!} \right) dx$$

$$= 1 - C(c,r) + \frac{\pi_0 r^c}{(c-1)!} \int_0^t \mu e^{-(c-r)\mu x} dx$$

$$= 1 - C(c,r) + \frac{\pi_0 r^c}{(c-1)!} \frac{1 - e^{-(c-r)\mu t}}{c - r}$$

$$= 1 - C(c,r) + C(c,r)(1 - e^{-(c-r)\mu t}) = 1 - C(c,r) e^{-(1-\rho)c\mu t}.$$

En los dos últimos pasos utilizamos que $c - r = c\,(1 - \rho)$.

q tiene un punto (t=0) con probabilidad positiva: $P(q=0) = P(N < c)=1 - C(c, r)$.

Análogamente, podemos obtener la distribución de *w:*

$$F_w(t) = \begin{cases} 1 + \dfrac{r - c + 1 - C(c,r)}{c - 1 - r} e^{-\mu t} + \dfrac{C(c,r)}{c - 1 - r} e^{-(1-\rho)c\mu t}, & \text{si } r \neq c - 1 \\ 1 - [1 + C(c,r)\mu t] e^{-\mu t}, & \text{si } r = c - 1 \end{cases}$$

Obviamente, tomando c = 1, recuperamos las fórmulas del modelo *M/M/1.*

Ejemplo. En una pequeña oficina hay un escáner alquilado para uso de los empleados. Aunque los trabajos a realizar varían en longitud, el tiempo de servicio puede aproximarse a una distribución exponencial con tasa media de 10 trabajos/hora.

En las 8 horas de trabajo diario, las peticiones de uso del escáner llegan aleatoriamente con una tasa media de 5 trabajos/hora. El tiempo del personal se valora en 5 euros por hora.

Las quejas recibidas por los empleados sugieren buscar mejoras del sistema actual:

- Una posibilidad es alquilar un escáner como el actual, a un coste de 11 euros diarios.

- Otra posibilidad es quedarse sólo con un escáner más rápido, atendiendo 15 trabajos/hora, con un coste de alquiler de 20 euros diarios.

El coste medio total al día (C_T) es el coste de alquiler (C_A) más el coste medio por el tiempo perdido por los empleados (C_E). Estudiar la opción más aconsejable.

La situación actual corresponde a un modelo *M/M/1* con λ=5, μ=10 trabajos/hora, de donde ρ = 0.5. Como

$$\begin{aligned} C_E &= (\text{número horas diarias perdidas}) \times (5 \text{ euros/hora}) \\ &= (8 \text{ horas/día}) \times (5 \text{ trabajos/hora}) \times (W \text{ horas/trabajo}) \times (5 \text{ euros/hora}) \end{aligned}$$

Como

$$W = \frac{L}{\lambda} = \frac{\frac{\rho}{1-\rho}}{\lambda} = \frac{1}{5} = 0.2$$

entonces C_E es 40 euros/día, y $\boldsymbol{C_T = 40 + 11 = 51}$ **euros/día**.

La posibilidad del escáner rápido cambia el modelo anterior, al tener ahora μ = 15 trabajos/hora, de donde ρ = 1/3 y W = 0.1 horas/trabajo, dando lugar a C_E = 20. Como C_A = 20, $\boldsymbol{C_{T1} = 20 + 20 = 40}$ **euros/día**.

Si decidimos utilizar dos escáners como el actual, el modelo es *M/M/2*, de donde $\rho = \lambda/(c\mu) = 5/(2\times 10) = 0.25$ y

$$W = \frac{1}{\mu}\left(1 + \frac{C(c,r)}{c(1-\rho)}\right) = \frac{1}{10}\left(1 + \frac{C(2,5/10)}{2(1-0.25)}\right) = 0.106$$

Así,

$$C_E = (40 \text{ trabajos/día}) \times (0.106 \text{ horas/trabajo}) \times (5 \text{ euros/hora}) = 21.2 \text{ euros}$$

y C_{T2} **= 21. 2 + 2 × 11 = 43.2 euros/día**.

Alquilar dos escáners pero ubicándolos en diferentes lugares de la oficina de forma que la mitad de los trabajos llegaran a cada escáner. Es decir, se tendrían dos modelos *M/M/1*, cada uno con λ = 2.5, μ = 10 trabajos/hora y ρ = 2.5/10 = 0.25.

El tiempo medio en cada sistema sería *W* = 0.1/0.75 = 0.133 horas/trabajo. Luego, C_{T3} **= 2(13.33 + 11) = 48.66 euros/día**.

Por tanto, debemos elegir alquilar el escáner rápido, que conlleva menores costes y menor tiempo perdido en el sistema.

Ejemplo. Una compañía telefónica quiere diseñar un servicio de información de números de teléfono. Desea determinar cuántos operadores contratar para satisfacer los siguientes criterios de diseño:

1. El tiempo medio esperando ser atendido no debe sobrepasar 2 minutos;

2. El 90% de las llamadas deben esperar menos de 2 minutos a que comience el servicio.

El tiempo que utilizan los operadores en atender las llamadas sigue un modelo exponencial con un tiempo medio de 4 minutos.

Se espera que las llamadas lleguen aleatoriamente con una media de 40 llamadas por hora. Las llamadas que se producen cuando todos los operadores están ocupados quedan a la espera hasta que uno queda libre.

En este sistema *M/M/c*, las tasas son λ = 40 llamadas/hora = 2/3 llamadas/minuto, μ=0.25 llamadas/minuto, con intensidad de tráfico *r* = 8/3.

Para que exista solución de equilibrio, debe ser λ / *cμ* <1, es decir, *c* ≥ 3 operadores.

Los criterios de diseño establecen que $W_q \leq 2$ minutos y $P(q \leq 2 \text{ minutos}) \geq 0.9$.

Para tres operadores (*c*=3), sabiendo que *C* (3, 8/3) = 0.8205, obtenemos los siguientes valores:

$$\rho = \frac{\lambda}{c\mu} = \frac{4/6}{3 \times 0.25} = 0.888$$

$$W_q = \frac{L_q}{\lambda} = \frac{C(c,r)}{c\mu(1-\rho)} = \frac{0.8205}{3 \times 0.25 \times (1-0.888)} = 9.846$$

$$\begin{aligned} P(q \leq 2) &= 1 - C(c,r)e^{-(1-\rho)c\mu t} = \\ &= 1 - 0.8205 \times e^{-(1-0.888)\times 3 \times 0.25 \times 2} = \\ &= 0.3054 \end{aligned}$$

En la tabla siguiente mostramos los resultados para varios valores de *c*. La columna de W_q está expresada en minutos.

c	$C(c,r)$	π_0	W_q	$P(q \leq 2)$
3	0.8205	0.0288	9.846	0.3054
4	0.4025	0.0637	1.207	0.7933
5	0.1733	0.0719	0.297	0.9460
6	0.0665	0.0740	0.080	0.9874

Conforme aumentamos el número de operadores, disminuye la probabilidad $C(c,r)$ de encontrar todos los operadores ocupados y aumenta la probabilidad π_0 de que el sistema esté vacío.

Con 4 operadores se satisface el primer criterio, pero para satisfacer además el segundo hacen falta 5 operadores. En ese caso, el factor de utilización es $\rho = r/5 = 0.533$, que indica que en media cada operador permanece ocioso casi la mitad del tiempo.

Ése es el precio de un buen servicio que satisface las condiciones de diseño.

5. Modelo *M/M/*∞: infinitos servidores

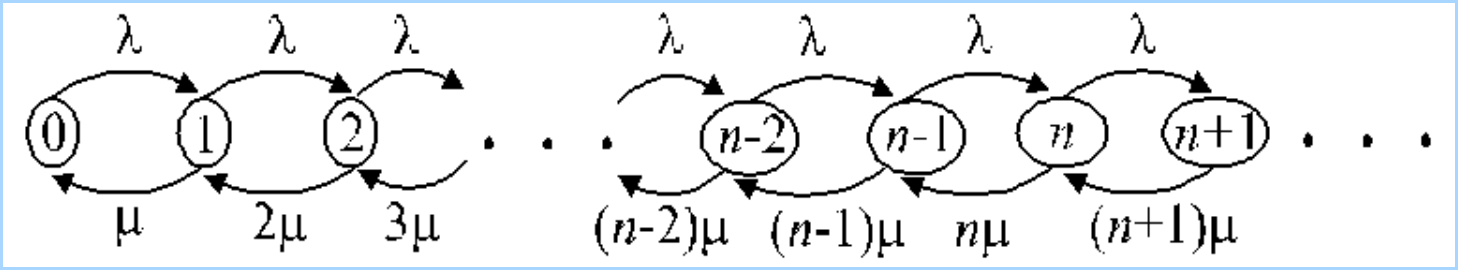
El sistema de espera tiene un número ilimitado de servidores, lo que significa que cada cliente que llega es servido inmediatamente.

A pesar de no haber competencia ni compartición de recursos, los resultados de este modelo pueden servirnos para estimar cantidades de interés en sistemas con un número *c* suficientemente grande de servidores.

Las tasas de nacimiento y muerte son

$$\lambda_n = \lambda, \quad n = 0, 1, 2, \ldots$$
$$\mu_n = n\mu, \quad n = 1, 2, 3, \ldots$$

y su diagrama de transición es



$$\begin{aligned}
\pi_0\lambda &= \pi_1\mu \\
\pi_1(\lambda+\mu) &= \pi_0\lambda + 2\mu\pi_2 \\
\pi_2(\lambda+2\mu) &= \pi_1\lambda + 3\mu\pi_3 \\
&\dots \\
\pi_n(\lambda+n\mu) &= \pi_{n-1}\lambda + (n+1)\mu\pi_{n+1} \\
&\dots \\
&\sum \pi_i = 1
\end{aligned}$$

$$\begin{aligned}
\pi_0\lambda &= \pi_1\mu \\
\pi_1\lambda &= 2\mu\pi_2 \\
\pi_2\lambda &= 3\mu\pi_3 \\
&\dots \\
\pi_n\lambda &= (n+1)\mu\pi_{n+1} \\
&\dots \\
&\sum \pi_i = 1
\end{aligned}$$

$$\begin{aligned}
\pi_1 &= \left(\frac{\lambda}{\mu}\right)\pi_0 \\
\pi_2 &= \frac{1}{2!}\left(\frac{\lambda}{\mu}\right)^2\pi_0 \\
\pi_3 &= \frac{1}{3!}\left(\frac{\lambda}{\mu}\right)^3\pi_0 \\
&\dots \\
\pi_n &= \frac{1}{n!}\left(\frac{\lambda}{\mu}\right)^n\pi_0 \\
&\dots \\
&\sum \pi_i = 1
\end{aligned}$$

$$\pi_0 = \left[\sum_{n=0}^{\infty}\frac{1}{n!}\left(\frac{\lambda}{\mu}\right)^n\right]^{-1} = e^{-\frac{\lambda}{\mu}}$$

$$\pi_n = \frac{(\lambda/\mu)^n}{n!}e^{-\frac{\lambda}{\mu}}, \quad n = 0, 1, \dots$$

Se obtiene que la v.a. N sigue una distribución de Poisson de parámetro $r = \lambda/\mu$, con la condición de estabilidad $r < 1$.

Por tanto, $L = \lambda/\mu$, que indica el número medio de servidores ocupados. Además, $\sigma_N^2 = \lambda/\mu$.

Como no hay cola, $L_q = W_q = 0$.

El tiempo medio en el sistema es el tiempo medio de servicio: $W = W_s = 1/\mu$ (también deducible del resultado de Little $W = L/\lambda$).

Más aún, la distribución de w es como la de s, exponencial de parámetro μ.

Ejercicios. Modelo de colas I

Al supercomputador de un centro de cálculo llegan usuarios según un proceso de Poisson con tasa 5 usuarios cada hora. Sabiendo que éstos consumen un tiempo de cómputo aleatorio cuya distribución puede suponerse exponencial de media 10 minutos y que la disciplina de cola es una FIFO, se pide:

a) El número medio de usuarios en el supercomputador y esperando para poder utilizarlo.
b) Suponiendo que hay usuarios esperando, obtenga el tamaño medio de la cola.

c) Si en la sala de espera hay 4 sillas ¿cuál es la probabilidad de que un usuario que llega a la sala tenga que esperar de pie?
d) ¿Cuántas sillas se necesitarían para que un usuario al llegar al sistema tenga una probabilidad menor del 10% de esperar de pie?

e) ¿Qué porcentaje de usuarios que llegan al servidor lo encuentran desocupado?
f) Obtenga el tiempo medio de los usuarios en el sistema y en la cola del mismo.

g) Obtenga la probabilidad de que un usuario espere más de una hora.

Ejercicios. Modelo de colas II

El tráfico en un centro de computación de mensajes, para una de las líneas de salida, llega según un patrón aleatorio de Poisson con un promedio de 240 mensajes por minuto. La línea tiene una velocidad de transmisión de 800 octetos por segundo. La longitud del mensaje es aleatoria con distribución aproximadamente exponencial con longitud media de 176 octetos. Se pide:

a) Calcular las medidas estadísticas de las prestaciones del sistema desde el punto de vista del usuario suponiendo un número elevado de buffers para mensajes.

b) Suponiendo que se desea colocar solamente buffers para que la probabilidad de que todos estén llenos en un determinado instante sea menor que 0.005, ¿cuántos hay que colocar? Calcular los estadísticos de las prestaciones desde el punto de vista del usuario para esta nueva situación.